R³Net: Efficient On-line Video Object Detection with pRoposal, tRacking, and Refinement

Taeyoung Kong
Department of Electrical Engineering

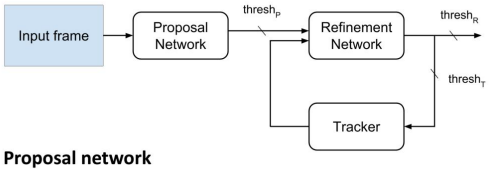
Motivation

- Video is an important data source for real-world vision tasks – e.g. autonomous driving, surveillance camera
- Processing video data is compute-intensive:
  - $50 camera : generate 1080p video stream at 25fps.
  - $1,000 Maxwell Titan X: run Faster R-CNN at 5fps
- Video has temporal and spatial locality
  - Exploiting locality, we can improve both accuracy and computation efficiency

Dataset

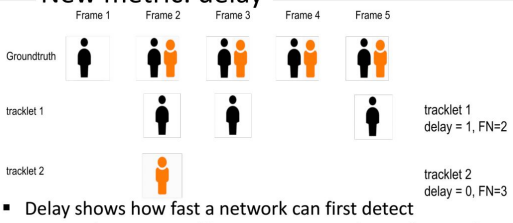
- **KITTI**
  - autonomous driving dataset: detection & tracking
- **Training**
  - Pre-trained on ImageNet and MS-COCO detection data
  - KITTI detection dataset for fine-tuning
  - KITTI detection is still image (not video) dataset with size of 1382 X 512 pixels.
- **Testing**
  - KITTI tracking dataset (21 video sequences) for testing
  - Each sequence has between 100~1000 frames
  - Each frame size is 1382 X 512 pixels
    - Easy: size < 40 Px. or Fully visible or truncation < 15%
    - Moderate: size < 25 Px. or Partly occluded or truncation < 30%
    - Hard: size < 25 Px. or Difficult to see or truncation < 50%

Framework



- **Proposal network**
  - Cheap/shallow networks – ResNet-10a, ResNet-10b
  - Detects newly emerging objects and feed into a refinement network
- **Tracker**
  - Estimate future locations of objects and feed into a refinement network
- **Refinement network**
  - Expensive/deep networks – ResNet-50
  - Only runs on partial of input image -> reduce computation
  - Calibrate detections proposed by a proposal network or a tracker

New metric: delay



- Delay shows how fast a network can first detect
- Once it detects a tracklet, it can recover from afterward miss

Results

model	mAP@hard	mAP@moderate	mAP@easy	FLOPS
Single model (ResNet-50)	75.1%	81.4%	88.0%	$\mathbf{256.0 \times 10^9}$
Ours (ResNet-10b + ResNet-50)	75.1%	82.0%	89.1%	$\mathbf{52.7 \times 10^9}$

Table. mAP comparison between a single model and our system on KITTI dataset

- On KITTI, R³Net is 5X faster than a single model
- It shows no loss of accuracy for hard objects, and even increase in accuracy for moderate/easy objects

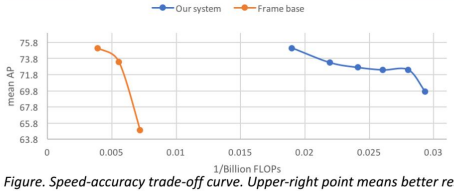


Figure. Speed-accuracy trade-off curve. Upper-right point means better result

- Our system pushes forward the performance envelope (speed-accuracy trade-off), on video object detection

model	car@moderate	pedestrian@moderate
Single model (ResNet-50)	1.4	5.000
Ours (ResNet-10b + ResNet-50)	1.368	4.182
Ours (ResNet-10a + ResNet-50)	1.246	4.273

Table. Result of experiments on delay. Delay is measured where precision is 0.8

- Our system shows better delay on both car and pedestrian objects with moderate difficulty

Conclusions

- R³Net proposes a framework for video object detection, composed of a proposal, a tracker, and a refinement
- R³Net achieves significant gains in computational time with no loss in accuracy on KITTI dataset
- We propose a novel evaluation metric for video object detection, delay, which shows how fast a detector can first detect a tracklet